Positive Sample Only Learning (PSOL) for Predicting RNA Genes in *E. coli*

Richard F. Meraz[1*], Xiaofeng He[2*], Chris H.Q. Ding[2] and Stephen R. Holbrook[1]
Physical Biosciences Division[1] and Computational Research Division[2]
Lawrence Berkeley National Laboratory, Berkeley, CA
rfmeraz,xhe,chqding,srholbrook@lbl.gov

Abstract

RNA genes lack most of the signals used for protein gene identification. A major shortcoming of previous discriminative methods to distinguish functional RNA (fRNA) genes from other non-coding genomic sequences is that only positive examples of fRNAs are known; there are no confirmed negatives – only intergenic sequences that may be positive or negative. To address this problem we developed the "Positive Sample Only Learning" (PSOL) method. This method can identify the most likely negative examples from an unlabeled set and is therefore able to distinguish putative functional RNA genes from other non-coding sequence. We compare RNA gene predictions using the PSOL method with previous large-scale analyses of the E. coli K12 genome.

1. Introduction

The importance of non-coding, or functional RNA (ncRNA/fRNA) in bacteria is unquestioned. Beside its canonical roles, RNA functions in a variety of ways as a regulator during gene expression [4, 5]. To date there are over fifty types of functional RNA molecules verified in *E. coli* [6] and several hundred other stable RNAs of unknown function identified by microarray analysis [10]. The development of automated computational methods for finding fRNA genes in genomic sequence has proven difficult because, unlike protein genes, RNA genes have few intrinsic features that can be exploited as signal in gene prediction algorithms. For example, there is a confounding lack of agreement among computational and experimental screens for RNA genes in *E. coli*. Less than ten percent of the several hundred RNA genes predicted by previous screens [1, 2, 9, 10] are mutually predicted by each method [7]. While experimental methods have had great success identifying a diverse class of functionally active RNA molecules, it has become clear that ncRNAs have varied stability and are expressed under such a variety of environmental and physiological conditions that predictions from an experimental method are unlikely to saturate any genome [3].

A two-class discriminative method is inappropriate for determining fRNA genes because only positive examples of fRNA genes are known. All negative training examples must be drawn from intergenic sequence which is certain to contain a – possibly significant – complement of undiscovered fRNA genes. To this end, we developed an algorithm for Positive Sample Only Learning (PSOL) in which a large margin classifier is iteratively trained to identify those intergenic sequences that are most likely to be true negative examples. In this paper we will sketch the PSOL method and describe a bioinformatics comparison of novel fRNA genes in *E. coli* predicted using PSOL and previous computational assays on the same organism.

2. Positive Sample Only Learning (PSOL)

Below we refer to parametrized sequences as "points", "examples", or "samples" in a sequence parameter space. The positive set, P, refers to known fRNA genes and the unlabeled set, U, refers to intergenic sequence. The negative set, N, contains sequences that are not predicted to be fRNA genes. The basic idea is to select those points from the unlabeled set which are most likely to be true negatives by first making an initial search of the set for examples maximally distant from the positive set and mutually distant from each other; and then iteratively adding to the negative set using a series of large margin classifiers. The points remaining in the unlabeled set after completion of the procedure are putative fRNA genes.

L1. Initial selection of negative set

Algorithm L1A. To form the initial negative set N we select m seed points from U that are most dissimilar from P using the point set distance, $d(x_i, P) = \min_{x_i \in P} \|x_i - x_j\|$,

∗ These authors contributed equally to this work.

such that the following is maximized:

$$\max_{N \subset U} d(N,P), \text{ where } d(N,P) = \sum_{x_i \in N} d(x_i,P) \quad (1)$$

This optimization is trivially solved by picking the $|N|$ points with the largest distance $d(x_i,P)$.

The set N may have many members close to each other. From the viewpoint of learning, there is a redundancy in N. Therefore, we maximize the distance between members of N:

$$\max_{N \subset U} d(N,N), \text{ where } d(N,N) = \sum_{x_i,x_j \in N} d(x_i,x_j) \quad (2)$$

The following optimization satisfies these two criteria:

$$\max_{N \subset U} [d(N,N)d(N,P)] \quad (3)$$

The exact solution of Eq. 3, however, is NP-hard. We propose the following approximate algorithm.

Algorithm L1B . The first point is chosen according to Eq 1. The rest of N is chosen incrementally. Suppose we already have several points in N. A new point i is selected based on maximum dissimilarity to the postive set,

$$\max_{x_i \in (U-N)} d(x_i,P) \quad (4)$$

and the maximum distance to the current set,

$$\max_{x_i \in (U-N),\ x_j \in N} d(x_i,x_j) \quad (5)$$

Now Eq. 4 is an exact solution to Eq. 1 and Eq. 5 is an approximate solution to Eq. 2. As in Eq. 3 these two criteria are combined into one:

$$\max_{x_i \in (U-N)} [d(x_i,P) \sum_{x_j \in N} d(x_i,x_j)] \quad (6)$$

Once the specified size of N is reached, the algorithm terminates and we set the initial negative training set as $N_{train} = N$.

L2. Negative set expansion

Given the current negative training set N_{train} and the current unlabeled set U, we perform negative set expansion. A Support Vector Machine (SVM) is trained on the dataset $P + N_{train}$ to obtain a large-margin decision boundary [11]. For this SVM, denote the support vectors in N_{train} as N_{SV}. Determine an SVM decision value $f(x_i)$ for each point in U. Using a threshold h, we select points in U:

$$N_h = \{x_i | f(x_i) \leq -h,\ h > 0\}$$

This is likely to produce unbalanced positive and negative sets for the next SVM training since often $|U| \gg |P|$ and hence the number of predicted negative examples could be large in each expansion. To balance the positive and negative sets we limit the size of newly predicted negative samples, N_{pred}, to be $r|P|$, where r is chosen beforehand. Assume the decision values are sorted in increasing order, then

$$N_{pred} = \{x_i | i \leq r|P| \text{ and } f(x_i) \leq -h\}$$

The total negative set becomes $N_{neg} = N_{neg} \cup N_{pred}$; the unlabeled set becomes $U = U - N_{pred}$; and the current negative training set becomes $N_{train} = N_{pred} + N_{SV}$.

L3. Stopping criteria of negative expansion

Negative expansion is repeated until (a) convergence or (b) the size of the current unlabeled set is equal to the predetermined number m of potentially positive examples, i.e. $|U| = m$.

3. Parameterization of genomic sequence and finding RNA genes

We put together a database of experimentally verified RNA genes in *E. coli* K12 using the RNA Family Database (RFAM) [6]. The m54 annotation of the genome and the database of RNA genes were used to annotate and extract all intergenic regions in the genome. All sequences were transcribed into RNA and cut into two sets of overlapping windows, both of length 80 nucleotides (nt.) with overlaps of 40 nt. and 20 nt., respectively. Each window was transformed into a vector consisting of sequence statistics (single, di, and tri nucleotide composition), the ΔG of folding as calculated by the Vienna RNA package [12], and similarity measurements between related genomes (WU-BLASTN W=3, available from http://blast.wustl.edu).

PSOL is performed on the feature vectors for two sets of windows. The sequence windows with 40 nt. overlap serve as "seed" windows such that predicted adjacent or overlapping windows are assembled into contiguous regions. These seed predictions are then extended and/or merged using the predicted 20 nt. overlap sequence windows. The resulting regions are putative RNA gene predictions which are used for further analysis.

4. Bioinformatics comparison

To evaluate PSOL, we measured recovery of five-fold hold-outs of the positive set. This was repeated five times. The percentage recoveries were: 84.57 ± 3.81, 85.17 ± 3.38, 84.79 ± 3.47, 88.56 ± 3.06, and 85.59 ± 3.04. Thus, PSOL shows good sensitivity. Because there is no trusted true negative set, we cannot estimate specificity. Intuitively, this can

be controlled by decreasing the number of remaining unlabeled examples that terminates the algorithm. In these experiments, we chose to terminate the algorithm when 900 examples remained. This yielded approximately the same coverage of the genome as previous computational assays (Table 1).

We compared the putative RNA genes predicted by PSOL to three previous large-scale attempts to find RNA genes in *E. coli* that each yielded a large number of unverified predictions. The methods used were: (1) pair stochastic context free grammars for modeling patterns of covariation in sequence alignments from related genomes (QRNA) [9], (2) profile based methods for annotating promoter and terminator regions bounding conserved genomic sequence (IBIS) [2], and (3) a whole genome microarray analysis of stable transcripts (Affymetrix) [10]. Table 1 summarizes the results from each method. Clearly, predictions of each screen are mostly mutually exclusive.

Experiment	*# discrete*	*% genome*	*% unique*
PSOL	381	0.88	0.59
Affymetrix	334	0.93	0.58
QRNA	275	0.89	0.57
IBIS	227	1.04	0.74

Table 1. Summary of RNA Genes predicted by recent experiments. Percentage genome reflects total nucleotides predicted normalized to genome length. Unique predictions do not overlap regions predicted by the other experiments and are also normalized to genome length.

To determine if the overlap between predictions of PSOL and each of the other methods could be accounted for by chance, we replaced PSOL with a random classifier ($P = 0.5$ for selecting any example with an increased probability, $P + 0.1$, that adjacent or cross strand predictions will occur once a given example is selected). We ran 500 simulations of the classifier and used the total nucleotide overlap between the simulated predictions and each of QRNA, Affymetrix, and IBIS as random variables for pair-wise comparisons. Only PSOL/QRNA with a $P << 0.05$ under a gaussian fit showed a significant overlap. Both QRNA and PSOL use comparative genomic data and implicit information about secondary structure as features for gene prediction which could account for this result.

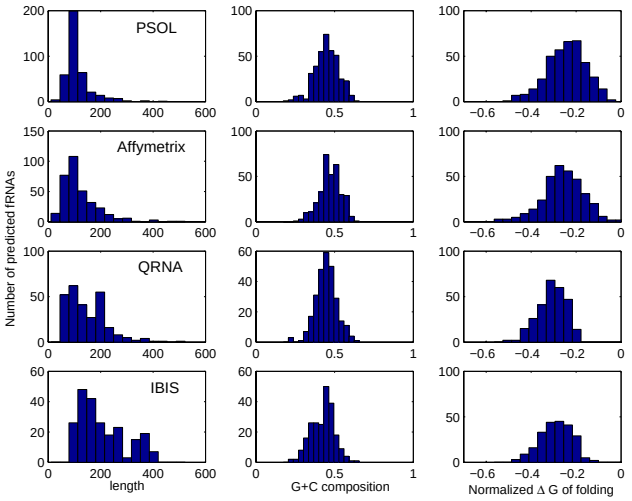

Figure 1. Histograms showing the distribution of lengths, G+C composition, and ΔG for putative functional RNAs.

We compared relevant global properties of RNA molecules: length, G+C compositon and normalized free energy of folding from the different genefinding methods. Several trends emerge. First, the predictions of both QRNA and IBIS are more stable than Affymetrix and PSOL. QRNA specifically identifies patterns of covariation in alignments that are likely under a model of RNA secondary structure; so a bias toward more stable secondary structures is not unexpected. PSOL, in contrast, uses additional sequence level statistics that could identify non-secondary-structure dependent features in RNA genes. Further pattern discovery, such as the application of clustering methods based on metrics defined in terms of secondary structure topology, will determine whether these differences in stability reflect different classes of RNAs being predicted by the various methods or merely biases in the methodologies. Second, G+C composition is similar for all methods with the exception of a small pool of low G+C compositon predicted by PSOL. Finally, the length distributions reflect shorter RNAs for QRNA, Affymetrix, and PSOL, and much longer predictions for IBIS. This difference is likely an artifact of the way promoter and terminator pairs are assigned in the IBIS procedure.

We conducted a sequence similarity search querying the predicted RNA genes against a database of all fully sequenced microbial genomes from NCBI-Genbank (151 genomes in May 2004). The results are shown in Figure 2. All of the methods have hits distributed throughout the genomes – the particularly dense areas correspond to in-

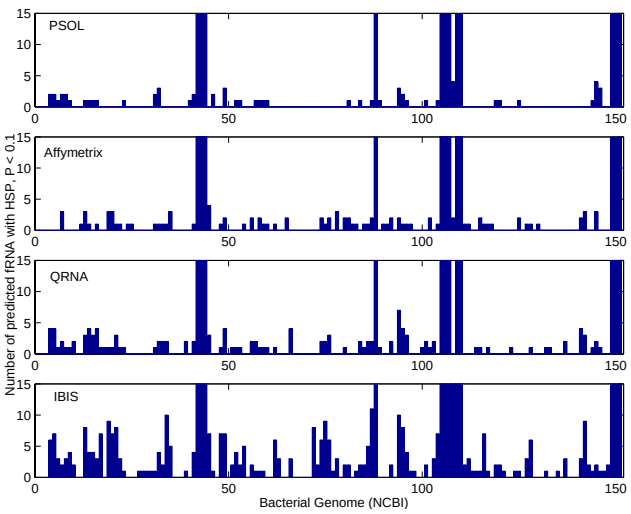

Figure 2. High Scoring Pairs of predicted fRNA genes distributed in 151 fully-sequenced microbial genomes from Genbank. (WU-Blast P<0.1 W=4)

stances of closely related *E. coli*, *Salmonella*, *Shigella*, and *Yersinia* species. Many of the IBIS predictions encompass low-complexity genomic sequence such as repeat regions. Therefore, many of their hits to more distant genomes are likely spurious. While a BLAST comparison can accurately identify significantly similar RNA molecules, it does not consider the secondary structure constraints necessary for identification of remote RNA homologs. The next stage of our analysis will employ a recently developed secondary structure guided alignment search [8] and newly developed kernel-based methods for secondary structure comparison (Karklin Y., Meraz R.F., and Holbrook S.R, unpublished results) to determine predictions that warrant experimental study.

5. Conclusion

Positive Sample Only Learning can be applied to any classification problem in biology where examples from only one class are known with certainty. The current application to RNA genefinding shows good sensitivity. Although we currently lack experimental validation, the predicted RNA genes have a significant overlap with predictions of QRNA and also have – possibly incidental – overlap with uncharacterized transcripts from a microarray experiment. Predictions show calculated free energies of folding and length distributions that are consistent with those observed for known small RNA genes. Importantly, each of the methods predict RNA genes with significant sequence similarity to a small group of related genomes. Further bioinformatic (e.g. remote homolog charaterization using secondary structure information) and experiemental analysis are necessary to determine which predictions should be annotated as true functional RNA molecules.

6. Acknowledgements

This work was supported by grant 5RO1H6002665-02 from the NHGRI. R. Meraz thanks E. Frise, and E. Smith for use of the excellent computational resources of the Berkeley Drosophila Genome Project.

References

[1] R. J. Carter, I. Dubchak, and S. R. Holbrook. A computational approach to identify genes for functional RNAs in genomic sequences. *Nucleic Acids Res*, 29(19):3928–38, 2001.

[2] S. Chen, E. A. Lesnik, T. A. Hall, R. Sampath, R. H. Griffey, D. J. Ecker, and L. B. Blyn. A bioinformatics based approach to discover small RNA genes in the Escherichia coli genome. *Biosystems*, 65(2-3):157–77, 2002.

[3] S. R. Eddy. Noncoding RNA genes. *Curr Opin Genet Dev*, 9(6):695–9, 1999.

[4] V. A. Erdmann, M. Z. Barciszewska, M. Szymanski, A. Hochberg, N. de Groot, and J. Barciszewski. The non-coding RNAs as riboregulators. *Nucleic Acids Res*, 29(1):189–93, 2001.

[5] S. Gottesman, G. Storz, C. Rosenow, N. Majdalani, F. Repoila, and K. M. Wassarman. Small RNA regulators of translation: mechanisms of action and approaches for identifying new small RNAs. *Cold Spring Harb Symp Quant Biol*, 66:353–62, 2001.

[6] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy. Rfam: an RNA family database. *Nucleic Acids Res*, 31(1):439–41, 2003.

[7] R. Hershberg, S. Altuvia, and H. Margalit. A survey of small RNA-encoding genes in Escherichia coli. *Nucleic Acids Res*, 31(7):1813–20, 2003.

[8] R. J. Klein and S. R. Eddy. RSEARCH: Finding homologs of single structured RNA sequences. *BMC Bioinformatics*, 4(1):44, 2003.

[9] E. Rivas and S. R. Eddy. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2(1):8, 2001.

[10] B. Tjaden, R. M. Saxena, S. Stolyar, D. R. Haynor, E. Kolker, and C. Rosenow. Transcriptome analysis of Escherichia coli using high-density oligonucleotide probe arrays. *Nucleic Acids Res*, 30(17):3732–8, 2002.

[11] V. N. Vapnik. *Statistical learning theory*. Adaptive and learning systems for signal processing, communications, and control. Wiley, New York, 1998.

[12] S. Wuchty, W. Fontana, I. L. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49(2):145–65, 1999.